Question

1. Why we need to do data association in tracking module?

- They are using an object detector to provide the Kalman Filter with N measurements. As the detector results can be noisy, they need to design a data association mechanism to decide which detection to pair with a predicted object state and which detections to treat as outliers.

2. Which method does this module use for the data association? How it works?

- They adopt the fairly standard practice of using the Mahalanobis distance instead. This distance m measures the difference between predicted detections $\mathrm{H}\hat{\mu}_{t+1}$ and actual detections o_{t+1} weighted by the uncertainty about the prediction as expressed through the innovation covariance S_{t+1} :
  - $\mathrm{m}=\sqrt{(o_{t+1}-\mathrm{H}\hat{\mu}_{t+1})^T S_{t+1}^{-1}(o_{t+1}-\mathrm{H}\hat{\mu}_{t+1})}$ (6)
- They also adopt the orientation correction approach from the AB3DMOT baseline. Specifically, when the angle difference between the detection and prediction is between 90 and 270 degrees, they rotate the prediction's angle by 180 degrees before calculating the Mahalanobis distance. A large angle difference like that usually stems from the detector that outputs an incorrect facing direction of the object. Furthermore, it is unlikely that the object makes such a large turn in the short time duration between consecutive frames. In their experiments, they show that the Mahalanobis distance provides better tracking performance than the 3D-IOU.

3. How does this module compute the covariance matrix for Kalman filter? Any other method to compute the covariance? What's the difference?

- Rather than using the identity matrices and heuristically chosen scalars to build the covariance matrices of the Kalman Filter as in AB3DMOT, they use the statistics of the training set data to estimate the initial state covariance, the process, and observation noise covariance.
- Specifically, their process noise models the unknown linear and angular accelerations. Therefore, they analyze the variance in the ground truth accelerations in the training data set. Let them denote the training set's ground-truth object center positions and rotation angles as ($x_t^{[m]}$, $y_t^{[m]}$, $z_t^{[m]}$, $a_t^{[m]}$) for timestamp t $\in$ $\{1\cdots T\}$ and object index m $\in$ $\{1\cdots M\}$. They model the process noise covariance as a diagonal matrix where each element is associated to the center positions and rotation angles ($Q_{xx}, Q_{yy}, Q_{zz}, Q_{aa}$) and estimated as follows:

$$\mathrm{Q}_{xx} = Var\left(\left(x_{t+1}^{[m]} - x_t^{[m]}\right) - \left(x_t^{[m]} - x_{t-1}^{[m]}\right)\right) \qquad (7)$$

$$\mathrm{Q}_{yy} = Var\left(\left(y_{t+1}^{[m]} - y_t^{[m]}\right) - \left(y_t^{[m]} - y_{t-1}^{[m]}\right)\right) \qquad (8)$$

$$Q_{zz} = Var\left(\left(z_{t+1}^{[m]} - z_t^{[m]}\right) - \left(z_t^{[m]} - z_{t-1}^{[m]}\right)\right) \qquad (9)$$

$$Q_{aa} = Var\left(\left(a_{t+1}^{[m]} - a_t^{[m]}\right) - \left(a_t^{[m]} - a_{t-1}^{[m]}\right)\right) \qquad (10)$$

- The above variances are calculated over $m \in \{1,...,M\}$ and $t \in \{2,...,T-1\}$. The Q's elements associate to the center velocity and angular velocity ($Q_{d_x d_x}, Q_{d_y d_y}, Q_{d_z d_z}, Q_{d_a d_a}$) are estimated in the same way as follows:

$$\left(Q_{d_x d_x}, Q_{d_y d_y}, Q_{d_z d_z}, Q_{d_a d_a}\right) = \left(Q_{xx}, Q_{yy}, Q_{zz}, Q_{aa}\right) \qquad (11)$$

- One might think that the above estimation seems to double count the acceleration. However, the above estimation is actually reasonable based on our process model definition. For example, consider the x component of the state and its velocity-related component d_x in the process model defined in Section 3.2:

$$\hat{x}_{t+1} = x_t + d_{x_t} + q_{x_t} \qquad (12)$$

$$\hat{d}_{x_{t+1}} = d_{x_t} + q_{d_{x_t}} \qquad (13)$$

- To estimate the two noise terms q_{x_t} and $q_{d_{x_t}}$, we have:

$$q_{x_t} = \hat{x}_{t+1} - x_t - d_{x_t} \qquad (14)$$

$$q_{d_{x_t}} = \hat{d}_{x_{t+1}} - d_{x_t} \qquad (15)$$

- where the predicted state components $\hat{x}_{t+1}$ and $\hat{d}_{x_{t+1}}$ can be estimated using the ground-truth state components x_{t+1} and $d_{x_{t+1}}$. The velocity-related components $d_{x_{t+1}}$ and d_{x_t} can be approximated as $x_{t+1} - x_t$ and $x_t - x_{t-1}$ based on our state definition in Section 3.1. And we can derive the equations as follows:

$$\begin{aligned} q_{x_t} &\approx x_{t+1} - x_t - d_{x_t} \\ &\approx \left(x_{t+1} - x_t\right) - \left(x_t - x_{t-1}\right) \end{aligned}$$

$$\begin{aligned} q_{d_{x_t}} &\approx d_{x_{t+1}} - d_{x_t} \\ &\approx \left(x_{t+1} - x_t\right) - \left(x_t - x_{t-1}\right) \end{aligned}$$

- The above approximation explains why we use the variance of accelerations to estimate the process model noise covariance from equation 7 to 11.
- Additionally, including the acceleration noise in both prediction equations in 12 and 13 also adds robustness to the data association. On the contrary, only including the acceleration to the velocity prediction in equation 13 will underestimate the uncertainty when predicting the next position. Consider the case that there is a very large real acceleration in the current time step which could not be

accounted for in the previously estimated velocity. In this case, we will have large uncertainty in predicting the next velocity. But we will only have small uncertainty in predicting the next position if we do not include the acceleration noise to the position prediction equation. By adding this additional acceleration noise in position prediction, we increase the predicted uncertainty of position. And that is used within the Mahalanobis distance and therefore the data association becomes more generous for matching and more robust. Similar reasoning also applies to other state variables.

- And for the elements related to the length, width, height, and other non-diagonal elements in Q, we assume their variances to have value 0.
- Our observation noise models the error in the object detector. Therefore, we analyze the error variance between ground-truth object poses and detections in the training set to then choose the diagonals entries of Rand the initial state covariance Σ_0. For this, we first find the matching pairs of the detection bounding boxe sand the ground-truth by using the matching criteria that the 2D center distance is less than 2 meters. Given the matched pairs of the detections and the ground-truth ($D_t^{[k]}$, $G_t^{[k]}$) for timestamp t $\in$ $\{1 \cdots T\}$ andmatched pair index k $\in \{1 \cdots K\}$, where

$$D_t^{[k]} = (D_{x_t}^{[k]}, D_{y_t}^{[k]}, D_{z_t}^{[k]}, D_{a_t}^{[k]}, D_{l_t}^{[k]}, D_{w_t}^{[k]}, D_{h_t}^{[k]})$$

$$G_t^{[k]} = (G_{x_t}^{[k]}, G_{y_t}^{[k]}, G_{z_t}^{[k]}, G_{a_t}^{[k]}, G_{l_t}^{[k]}, G_{w_t}^{[k]}, G_{h_t}^{[k]})$$

- we estimate the elements of the observation noise covariance matrix R as follows:

$$R_{xx} = Var\left(D_{x_t}^{[k]} - G_{x_t}^{[k]}\right)$$

$$R_{yy} = Var\left(D_{y_t}^{[k]} - G_{y_t}^{[k]}\right)$$

$$R_{zz} = Var\left(D_{z_t}^{[k]} - G_{z_t}^{[k]}\right)$$

$$R_{aa} = Var\left(D_{a_t}^{[k]} - G_{a_t}^{[k]}\right)$$

$$R_{ll} = Var\left(D_{l_t}^{[k]} - G_{l_t}^{[k]}\right)$$

$$R_{ww} = Var\left(D_{w_t}^{[k]} - G_{w_t}^{[k]}\right)$$

$$R_{hh} = Var\left(D_{h_t}^{[k]} - G_{h_t}^{[k]}\right)$$

- The non-diagonal entries of R are all zero. We set Σ_0 = R as we initialize the multi-object tracker with the initial detection results.